1
Introdução

Bioinformática é uma área de pesquisa científica interdisciplinar que envolve a Ciência da Computação (aplicação de técnicas) e a Ciência Biológica (conceitos) para entender e organizar as informações associadas a moléculas, em grande escala, oferecendo ferramentas que apoiam o trabalho de pesquisa no que se refere à análise de sequência, estrutura e função molecular [Luscombe et al. 2001].

O avanço tecnológico apresentado nesses últimos anos tem proporcionado um aumento no poder computacional e no desenvolvimento de novas e mais eficientes técnicas de processamento e pesquisa. Especificamente na área da biologia, a bioinformática acelerou o processo de estudo do material genético de inúmeros organismos, revelando, por exemplo, que a análise sistemática de todo o conteúdo genético de um organismo tem o potencial de levar à compreensão integral da genética, da bioquímica, da fisiologia e da patogênese dos micro-organismos [Brosch et al., 2001]. Um exemplo do crescente número de dados biológicos que se têm produzido é o GenBank [Benson et al. 2007], uma base de dados de nucleotídeos e proteínas construída pelo Centro Nacional para Informação Biotecnológica (NCBI). Dados estatísticos mostram que o tamanho do GenBank dobra a cada 18 meses [Benson et al. 2007].

Se por um lado a bioinformática acelerou o processo de geração de dados biológicos, por outro, ela originou alguns desafios. Além da grande quantidade de dados gerados, há muitos tipos de dados e cada um tem sua própria comunidade e seus próprios repositórios. Cada nova área de pesquisa que surge desenvolve sua própria representação dos dados considerando seus próprios conceitos. Para se ter uma ideia da heterogeneidade dos dados, existem mais de 231 diferentes bases de dados de vias metabólicas (*pathways*) [Goble and Stevens 2008].

Para a biologia molecular, parece que é mais fácil, mais desejável, ou mais conveniente criar novamente um Banco de Dados (BD) do que adaptar ou reutilizar recursos existentes. A natureza autônoma de muitos gestores de dados biológicos, somada com a volatilidade dos dados e a rápida evolução de

experimentos da ciência, leva a uma tendência de surgir recursos de dados múltiplos, altamente heterogêneos [Goble and Stevens 2008]. Além disso, existe o agravante de que esses dados biológicos são armazenados, na sua maioria, em arquivos no formato texto.

Observando este cenário, surgiram algumas questões importantes. Por que não usar um Sistema de Gerência de Banco de Dados (SGBDs)? Como integrar e persistir dados biológicos? E, como manipular os dados biológicos de forma eficiente? Embora a utilização de SGBDs para armazenamento e manipulação de dados biológicos pareça algo natural, por se tratar de um programa que utiliza estruturas de armazenamento e manipulação de dados bem definidas, técnicas de transação, controle de concorrência e acesso distribuído consolidadas, seu uso não é uma unanimidade no meio científico [Seltzer 2008] [Topaloglou 2004] [Jagadish and Olken 2004]. Existem dúvidas, porém, quanto à capacidade de representação, manipulação e armazenamento de dados biológicos em SGBDs comerciais.

Alguns trabalhos na literatura vêm investigando o uso de SGBDs para armazenamento de dados biológicos. *ProteinWorldDB* [Otto et al. 2010], *TcruziDB* [Luchtan et al. 2004] e *CryptoDB* [Puiu et al. 2004] são exemplos de projetos que propõem esquemas de dados genômicos utilizando SGBDs para gerência dos dados. Relacionado à integração de dados, existe uma grande variedade de técnicas, tecnologias e sistemas propostos. SBRML [Dada et al. 2010], BioDBnet [Mudunuri et al. 2009], Bio2RDF [Belleau et al. 2008] e [Smedley et al. 2008] são algumas propostas de mecanismos de integração de dados, que podem ser: arquiteturas orientadas a serviços, integração de link, *data warehousing*, integração de visão, arquiteturas orientadas a modelo, *workflows* e *mashups*. Já a manipulação de dados biológicos de forma eficiente, segue como uma oportunidade de pesquisa.

1.1. Caracterização do Problema

Um dos problemas relevantes em aberto diz respeito à forma de armazenar e manipular dados biológicos. Informações inerentes ao domínio, tais como DNA, proteína, aminoácido, nucleotídeo e seus derivados e relações (dogma central da biologia molecular), são atualmente representadas como simples cadeias de caracteres sem nenhuma preocupação com o seu significado. Para acesso aos dados, algumas estruturas de índices para manipulação de sequências têm sido propostas, destacando-se a *suffix tree*

[Hunt et al. 2002] [Hunt et al. 2001] por ser uma estrutura de dados versátil e muito eficiente para a execução de consultas, podendo ser construída em tempo linear, caso consiga ser armazenada em memória principal [Cheung et al. 2005].

Contudo, não existe, na literatura, uma proposta para representar, armazenar e manipular, de forma adequada e apropriada, uma sequência biológica, ou ainda, informações derivadas, tais como a relação entre as sequencias que fazem parte do dogma central da biologia molecular e alinhamento entre proteínas. Ainda não há uma estrutura específica para o armazenamento e manipulação de dados biológicos. A maioria dos sistemas persiste os dados em arquivos no formato texto, e.g. programas da família BLAST [Altschul et al. 1990], SSEARCH [Pearson 1991] e BioParser [Catanho et al. 2006],e sistemas que utilizam SGBDs persistem a sequência em estruturas do tipo string ou BLOB (*Binary Large OBject, Basic Large OBject*), na sua forma original. Esta prática facilita a carga dos repositórios a partir de arquivos texto, porém o acesso aos dados é limitado aos operadores tradicionais [Lifschitz 2007].

Um grande ponto de discussão é com relação ao ganho de se investir em um SGBD para armazenar as sequências ao invés de arquivos textos, como já é feito pela maioria dos sistemas biológicos. Conforme já mencionado, entende-se como natural o uso de um SGBD para gerenciar grandes volumes de dados. No entanto, as estruturas de dados disponíveis pelos SGBDs convencionais não são adequadas para a gestão de sequências biológicas.

O problema de se tratar uma sequência biológica como uma simples palavra (*string*) ou BLOB, que é utilizado para armazenar qualquer dado, arquivos em geral, está na perda de informação semântica. Uma “string biológica” possui interpretações bem definidas, e.g. aminoácidos, proteínas, regiões codificadoras, etc., e características específicas que a difere de uma string de palavra, e.g. comparação e similaridade não são simples casamentos de padrões.

Independente da forma que se utilize para armazenar e gerenciar dados biológicos, arquivos texto ou tipos (*string*) ou BLOB no caso de SGBDs, não existem mecanismos apropriados para responder questões tais como: identificação de genes únicos e genes homólogos (parálogos e ortólogos).

1.2. Objetivos

Entendendo as dificuldades existentes na gestão de dados biológicos e identificando a falta de estruturas de dados específicas para representação, armazenamento e manipulação de sequências e seus derivados, propõe-se como Tese de Doutorado um modelo conceitual biológico para representar informações do dogma central da biologia molecular, bem como um tipo abstrato de dado (ADT – do inglês *Abstract Data Types*) específico para a manipulação de sequências biológicas e seus derivados.

Vale lembrar que a alternativa de um sistema dedicado já é adotada em outras áreas de conhecimento, como por exemplo, Sistemas de Informação Geográficas (SIG) [Dias et al. 2005] [Neteler and Mitásová 2008], e Banco de Dados Temporais [Simonetto and Ruiz 2000].

1.3. Estrutura do Trabalho

Este trabalho está estruturado como segue: no próximo Capítulo é apresentada a fundamentação teórica e trabalhos correlatos, trazendo uma visão geral dos assuntos relacionados a esta Tese. No Capítulo 3 é descrita a proposta de trabalho, compreendendo o modelo conceitual de dados biológico e mecanismos para acesso e manipulação de sequências biológicas. O Capítulo 4 trata da prototipação do que foi proposto nesta Tese, onde mostramos a expressividade do modelo conceitual (correto e robusto) através de uma implementação utilizando uma modelagem relacional e a realização de consultas que retornam informações de domínio biológico. Por fim, o Capítulo 5 está dedicado às conclusões e trabalhos futuros.